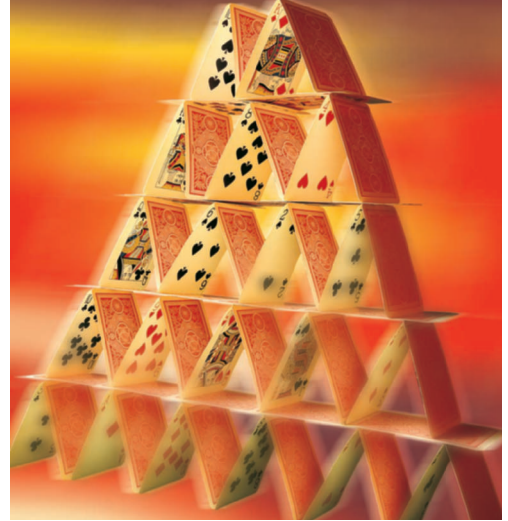


There's content everywhere, but not the information you need. Content analysis can organize a pile of text into a richly accessible repository.

Ramana Rao



From Unstructured Data to Actionable Intelligence

Businesses create huge amounts of potentially valuable content in the form of documents such as e-mail messages, drafts, project plans and reports, operational memos, customer reports, invention proposals, and research notes. However, organizations typically use such documents once and then lose them, despite the savings they could realize by reusing them. Internal documents might constitute companies' most ineffectively utilized asset today.

The problem afflicts individual knowledge workers as well as entire organizations. The individual typically suffers from what we call information overload—one poor soul, awash in rising tides of content and data. Statistics indicate that workers waste many hours searching for, sorting, and assessing information, incurring a significant organizational productivity cost. For example, International Data Corp. (IDC) estimates that an enterprise with 1,000 knowledge workers loses a minimum of \$6 million a year in the time workers spend searching for—and not finding—needed information (*The High Cost of Not Finding Information*, IDC, Apr. 2003, IDC #29127).

But there's an even larger concern, a long-term institutional problem. Workers usually don't waste their time fighting systems or tasks that don't pay off. People almost always move on when they can't find useful information quickly. And they

are unlikely to grind away at digesting poorly organized or apparently featureless piles of documents. Thus, organizations don't draw on reservoirs of information that could influence a particular decision, task, or project. Ultimately, this leads to uninformed decisions, overlooked risks, and lost opportunities.

Solving this problem requires an approach to organizing and cataloging content that is more active than current methods. In particular, using content effectively requires knowing more about the content—having access to information codified as document and collection metadata. Current systems often automatically capture various process-based metadata—for example, file system attributes such as author, title, size, creation date, and so on. Much more useful, and much rarer, is metadata about the document's actual content—for example, content summaries, topics covered, and people or companies mentioned. The "Unstructured—Not!" sidebar explores this topic further.

This article explains two key technologies for generating metadata about content—automatic categorization and information extraction. These technologies, and the applications that metadata makes possible, can transform an organization's reservoir of unstructured content into a well-organized repository of knowledge. With metadata available, a company's search system can move beyond simple dialogs to richer means of access that work in more situations. Information visualization, for example, uses metadata and our innate visual abilities to improve access. Besides

Inside

**Unstructured—
Not!**
**Out of the Box on
Search and Browse**

Unstructured—Not!

Neither content nor knowledge work is truly unstructured. Content, despite often being called “unstructured data,” is shaped—first, by intrinsic aspects of representation and expression and, second, by the social context in which it is produced and consumed.

Consider a physical magazine. You would hardly call it unstructured. You can riffle through it quickly, rattling off observations: Here’s the table of contents. Here’s the editor-in-chief. Oh, I hate that color, but what a nice dress. This is an ad, this is a feature article, and this one is really confused.

Our problem, as humans, starts not with one magazine, but with the stacks of unread magazines that pile up over months. Even worse are the piles of electronic documents on intranets and the Internet.

The term unstructured data refers to the difficulty of applying IT to routing and accessing content—slicing and dicing and manipulating it in all the ways typical of data stored in rows and columns in relational databases. The implied challenge is devising methods for extracting the latent structure embedded in content. Without access to such a structure, computers can’t assist us in dealing with our piling-up volumes of information.

Meanwhile, automation has focused on certain kinds of highly structured routine work that aligns well with databases. Yet, workplace anthropologists commonly point out that even after automation, these jobs involve much more “unstructure” than the boss knows or, perhaps, wants to know. Knowledge work, on the other hand, has much more structure in it than current tools support.

We access information for various purposes and in various ways according to our purpose. Sometimes we’re surveying an area of knowledge, trying to get a general understanding of what it’s about or what’s available. At other times we’re searching for specific answers.

Sometimes we wander with a vague sense that something important is on the verge of crystallizing in our understanding. Other times we’re skiing a series of tight turns, narrowing in on our target. It is this range of purpose and context that we can better address by providing a richer set of information access tools based on exploiting metadata.

better access, metadata enables intelligent switching in the content flows of various organizational processes—for example, making it possible to automatically route the right information to the right person. A third class of metadata applications involves mining text to extract features for analysis using the statistical approaches typically applied to structured data. For example, if you turn the text fields in a survey into data, you can then analyze the text

along with other data fields. All these metadata-powered applications can improve your company’s use of its information resources.

AUTOMATIC CATEGORIZATION

Portals and content management systems often claim to provide metadata, but they in fact rely on humans to provide any metadata beyond what they automatically capture as they store or route documents. Although this approach can work well in structured work flows, it is untenable in loosely structured knowledge work, especially as organizations focus on knowledge workers’ productivity.

A categorization system can generate collection catalogs or directories automatically.

A categorization system creates and maintains a hierarchical structure of categories—called a taxonomy—and assigns documents to the categories. A typical application is to use the taxonomy as a navigable directory for a high-value collection of private content—that is, you can create something like Yahoo for your own content.

Taxonomies typically blend classification schemes (such as the Dewey decimal system) and controlled vocabularies (such as Library of Congress subject headings) used in systems for cataloging and indexing library collections. They arrange subjects or topics into a hierarchy from general to specific categories. For example, you might see Art at the top, Postmodern Neon Art in the middle, and American Postmodern Neon Art in the 21st Century at the bottom.

A taxonomy’s effectiveness and design depend on many factors. It’s easy to get lost in a thicket of theoretical concerns, but remember that a taxonomy’s ultimate evaluation occurs at the point of use. Thus, who the users are, what they will be doing, and what they know and understand are much more important than abstract properties of knowledge.

General collections and tasks aren’t likely to be the highest-value applications for enterprise categorization in the near term. For example, would the pharmaceutical company Pfizer want its intranet to present its internal content in a news-oriented view of the world as would Reuters or Dow-Jones? Or would it do better to organize this content in a view of the world according to Pfizer? Perhaps it would ultimately want both, but the latter holds more obvious value. After all, a privileged understanding in a focused area of knowledge or practice is what creates a competitive advantage.

So, a general-purpose taxonomy would probably be less useful than appropriately specialized or even private taxonomies. Focused taxonomies are likely to make finer-grain discriminations within topics in more specialized



collections, and are also likely to better match the language and the purposes of specialized users and uses. For example, the Medical Subject Heading (MeSH) taxonomy, which has evolved and been used over many years in medical research, is more suited than a general taxonomy for organizing Pfizer's reports for its thousands of drug researchers.

Beyond defining and managing a taxonomy, a categorization system must accurately assign documents to one or more categories. Most of the leading products do this using a mixture of several methods, including linguistic analysis, statistical inference, machine learning, and rule-based processing. Debates about the relative value of the different methods and the best ways of blending them will likely continue. For the moment, other aspects besides core accuracy will probably make the critical difference in your choice of categorization system. Functionality for managing taxonomies across their full life cycle and integration of the categorization systems into enterprise environments are particularly important. These features can have a major impact on the total cost of ownership and the likelihood of success during deployment.

Once the categorization system has assigned a document to categories, the system can use the category tags or codes for several purposes:

- *End-user browsing.* As the analogy with Yahoo suggests, a hierarchical directory allows browsing while gaining some of the precision and control of searching. It effectively turns a search's command line interface into a hierarchical menu of topics. Browsing the taxonomy imparts an overall sense of what's available and how it's organized, and then allows drilling down to specific topics of interest.
- *Document routing.* Whereas the purpose of end user browsing is to let the user access a pool of past content, this application's purpose is to process the stream of new content and trigger appropriate actions or responses within the enterprise. The system can notify a user automatically as it assigns documents to categories that relate to that user's work. It can route documents to the appropriate employee for processing.
- *Enhanced search and browse.* A system can blend a search function with a taxonomy in several ways that enhance searching or enable a tighter interplay between searching and browsing. Showing search results organized by the categories of the matching documents helps a user more quickly digest the results. Search results can also become a point of departure for browsing, because a document's category will probably contain other relevant documents that might not have matched the query. A related technique that has proven popular with Internet

Out of the Box on Search and Browse



The two most common paradigms for finding documents are search and browse. With years of Internet use behind us, the two approaches' strengths and weaknesses are familiar: Search is precise yet brittle. Browse is robust but vague.

Searching works quite efficiently when it works at all. If you know what you want and how to describe it, and if you are looking in the right place, nothing works better. But how often do all these conditions coincide? Search is as precise as a scalpel or laser, but can it alone help you discover or understand the disease?

Meanwhile, browse has essentially the opposite profile. You can always browse, but it's not clear what you're doing. You scan a page, you read a little, you click, you're somewhere else, you aren't sure where you are, you lose track of time.

Rather than seeing search and browse as two alternatives, think of them as two ends of the spectrum. The structuring and access technologies described in this article extend, blend, or mix the best of search and browse in various ways.

Categorization by providing a nested hierarchy of queries from general to specific provides the ease of browsing while letting you get more and more precise, as with advanced search. Information extraction enables dialogs that offer menus of precise terms you might not know or remember. And visualization gives you a way of understanding thousands of categories or documents at a time, along with a way to navigate quickly to relevant content.

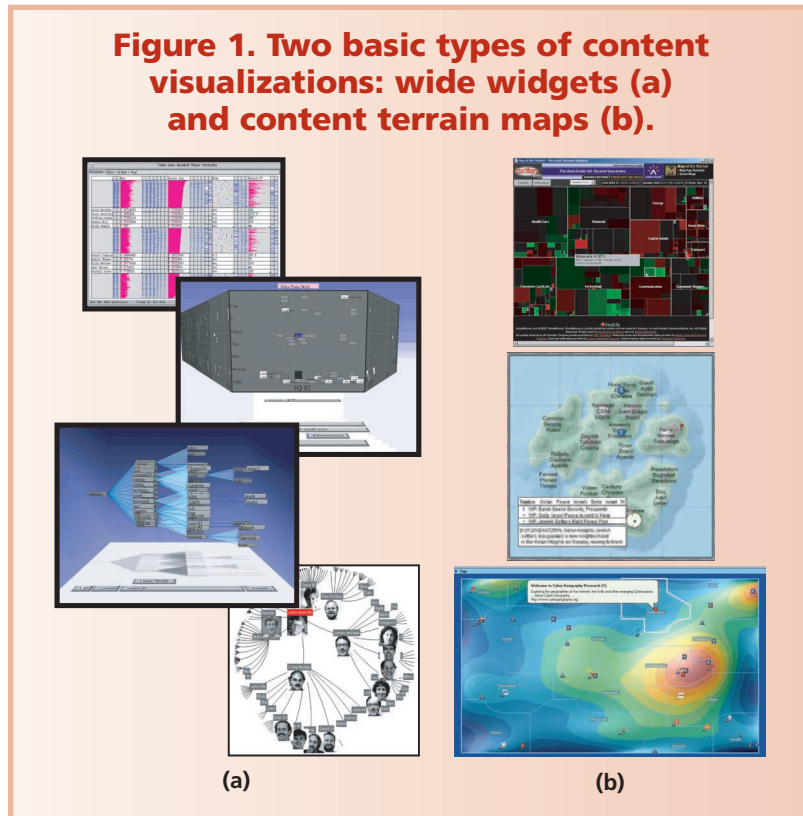
directories is to search for categories and then to navigate up and down from the matching categories. The "Out of the Box on Search and Browse" sidebar discusses the strengths and weaknesses of search and browse.

INFORMATION EXTRACTION

Information extraction systems extract elements of information from documents and collections. This technique resembles what we do as we scan an article to assess its relevance to our goals. Based on linguistic analysis and scanning for patterns among words and phrases, extraction identifies many of the things people can quickly find in text without reading for deeper meaning. Two levels of extraction are common:

- *Entity extraction* focuses on identifying "named entities" such as people, organizations, products, and places. Also important are other special noun groups including large noun phrases that indicate topics or concepts, as well as entities such as dates, quantities, and measurements. Even this basic level of identification can dramatically

Figure 1. Two basic types of content visualizations: wide widgets (a) and content terrain maps (b).



improve higher-level capabilities including categorization, search, and automatic summarization.

- **Fact extraction** spreads out from entities and topics, connecting and contextualizing them in relationships, thus expressing facts.

The simplest level of fact extraction is to determine the roles that various entities play and the relationships among them. For example, this type of extraction could determine that a particular person is the CEO of one company and also a board member for another company. This form of link creation lets us quickly use facts in documents as a way of understanding connections in the larger world.

Information extraction techniques can also link and contextualize topics in various ways. That a document concerns two topics together might be the key reason to select it. The fact that two topics appear more and more regularly together in documents might suggest a synthesis or the emergence of an important new perspective.

Another important class of facts to identify is significant events or event classes. For example, the announcement of a merger and acquisition, along with the roles played by organizations as acquirer and acquired, would be information of this type.

Extraction can provide a rich stream of additional tags that a content management system or database can store as metadata accompanying documents to support various

purposes. Further, fact extraction can enable more specialized forms of querying or action triggering:

- **Document preview.** Extracted entities and facts—when displayed in search results, directory entries, or other places that reference documents in user interfaces—can provide clues to a particular document’s usefulness to a specific task. By seeing a little bit more, but not too much, users can decide efficiently whether to examine the whole document or complete their task using just the preview.
- **Content packaging and routing.** The richer set of tags that extraction obtains can enable more sophisticated forms of packaging and routing. For example, these tags might let a publisher slice and dice its content into new channels or product offerings. This creates opportunities for new revenue streams and enhances the experiences of existing customers, increasing their loyalty and their revenue potential over time.
- **Surveillance.** Information extraction schemes can scan new documents for new events of interest. For example, a financial industry application might detect mergers and acquisitions or management personnel changes in new streams or Web sites. Security agencies could use information extraction to detect suspicious patterns of activity in intelligence information.
- **Answering questions.** We can often express our specific questions as queries structured over roles, relationships, and entities. Combining search with fact extraction on results sets provides a way of answering many types of questions using a document collection—or at least of zeroing in on the most relevant content.

INFORMATION VISUALIZATION

Information visualization involves using visual techniques to increase the bandwidth of our interaction with information. In many ways, you might think of visualization as supercharged browsing, but current Web browsing and directory navigation often require much more reading and mechanical effort than visualization would entail.

Visualization takes advantage of our evolved preattentive, automatic skills for processing large amounts of visual information and for staying oriented in space and retaining spatial information. For example, in an array of thousands of points, we can quickly spot a single red one. We can see small discontinuities or changes of texture or shade. Groups, parallel structures, and shapes pop out at us. And

we are quite good at remembering that a certain sentence was in the upper-left corner or how high it was located on the page. Taking advantage of these innate skills is the design opportunity of visualization.

Effective visualizations function like maps. Generally, maps let you survey or assess an entire territory and also navigate to specific locations. For example, a map would let you understand the neighborhoods and terrain of San Francisco and also plan a route from your hotel to the Moscone Center.

As with maps, different visualization techniques might emphasize one task or another. Some techniques emphasize big-picture thinking, whereas others might emphasize the detail and control necessary to navigate tightly or answer specific questions. Commercial applications are increasingly applying two types of visualizations to content collections; Figure 1 shows examples of both:

- *Content terrain maps* are directly analogous to geographic terrain maps. The visualization system generates representations of content sections and subsections, typically using automatic analysis to assign locations and area based on the properties of documents in the collection, and to render glyphs for subsections or items in the collection. This type of visualization supports various forms of interactions with the map surface and with extra tools that might search or highlight sections or items on the map. Typical examples are Smart Money's Map of the Market (<http://www.smartmoney.com/marketmap>) and Antarctica maps (<http://www.antarctica.net>).
- *Wide widgets* are like user interface objects that make up a graphical user interface, but they typically show many more objects at a time and provide much more effective spatial management. Usually, this type of visualization shows an interactive, visual structure that mirrors some central spine in the information. Typical examples are the Inxight Star Tree (<http://www.inxight.com/VizServerDemos/demo/orgchart.html>) and TheBrain (<http://www.thebrain.com>), which are tools for navigating large hierarchies and graphs.

Visualizations support several uses that synthesize searching and browsing:

- *Collection overview.* By showing a rendering of the entire collection or several levels of the taxonomy at once, a visualization can help a user quickly get an overall sense of the collection and its structure. The concreteness of an effective visual structure has much to offer as a mental map that can act as a resource for interaction over time.
- *Rapid navigation.* A visualization can tightly integrate navigation and interaction with pattern or anomaly observation. For example, if you regularly use Map of the Market, you can identify and easily access the information about the companies that stand out today or that represent a

group that stands out. Similarly, Star Tree lets users rapidly move up and down a taxonomy several levels at a time once they learn how the taxonomy is organized.

- *Interpreting items in context.* "Spotlighting" search results on a graphical visualization—using salient visual features to mark matching items, for example, marking matches with red dots or flags—is a powerful technique for helping users interpret search results. Spotlights let users quickly see clusters of matches and thus focus on the most relevant areas of the collection. Documents that are close to matches, but that didn't themselves match, might in fact be quite relevant. Conversely, users could either quickly dismiss isolated matches or, because of their unusual context, embrace them as the most interesting. This type of result marking provides a specific example of the more general power of visualizations to help users interpret local features in their context.

APPLICATIONS FOR UNSTRUCTURED DATA MANAGEMENT

Organizations can apply categorization, extraction, and visualization technologies in several ways to increase the efficiency of their content use. A typical deployment involves a number of infrastructure elements, as Figure 2 shows, that support both IT administrators and content organizers in the full process of deriving metadata from content and using the metadata in user-facing applications.

The most common applications help users find and understand documents as the need arises in their broader activities. In such cases, a user becomes explicitly aware of his or her information need and then searches or browses content sources to fulfill that need. Applying categorization, extraction, or visualization can lead to faster, better end user information access.

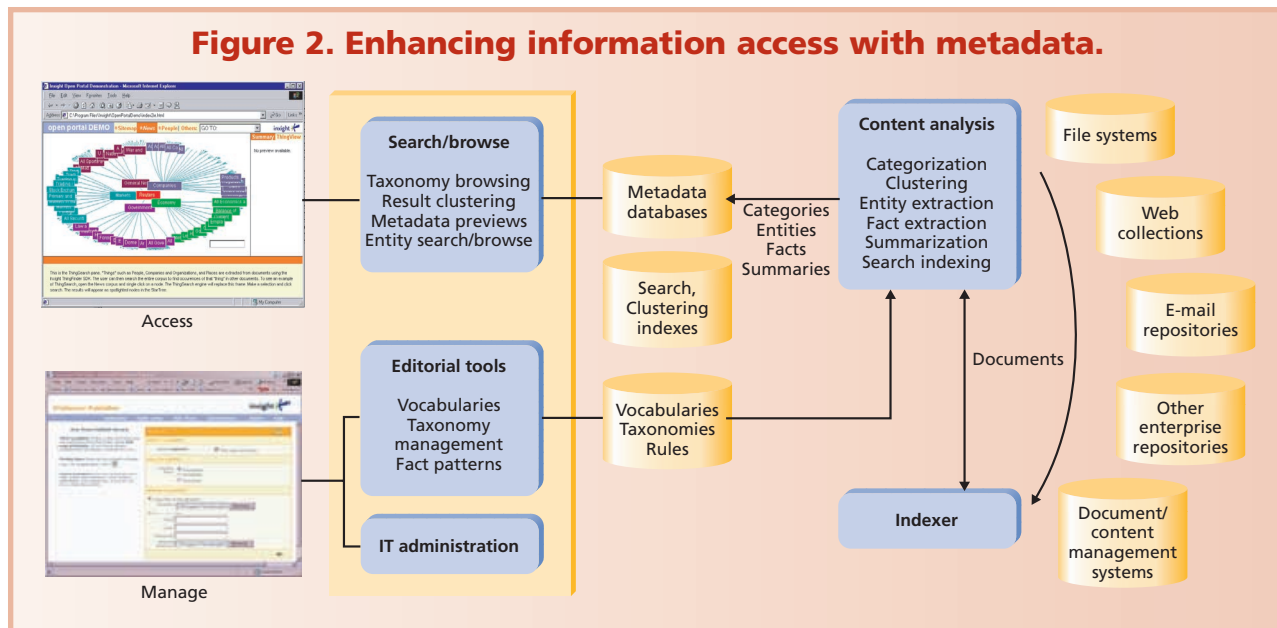
Two other general types of applications—routing and mining—leverage content-based metadata in broader organizational processes. Routing applications use the metadata to increase the degree of automation or intelligence in the process of moving content through the organization and its surrounding network. Mining applications enable the statistical analysis of content collections or flows to discover patterns or to drive organizational attention.

All three types of applications are becoming common in industry and market sectors. For an illustration of the range of potential applications and their organizational benefits, let's look at what the technologies I've described can offer three specific sectors—publishing, government intelligence, and life sciences research and development.

Publishing

In the publishing industry, content itself is the primary product (or the carrier of the service). Not surprisingly, publishing enterprises invest heavily in human activities and content technologies to create rich metadata that enables better access. Enhanced searching and browsing

Figure 2. Enhancing information access with metadata.



capabilities can generate greater customer usage.

Furthermore, content metadata supports the publishing industry's inherent routing-style processes for generating a broad range of content-based products and services. Timely, well-targeted content can have extremely high value for particular audiences and so represents significant revenue opportunities. Content packaging, repackaging, aggregation, syndication, and integration all have the purpose of more effectively reaching a greater number of customers and generating greater value.

The publishing industry is also a leading sector for mining-style opportunities. Many content customers are not interested in the content per se, but in what it reveals about, for instance, the market, trends, or broader realities. For example, a customer might purchase content to support competitive intelligence or corporate licensing activities. Needs such as these will also generate opportunities for publishers to provide content-mining applications or tools to their customers.

Intelligence and law enforcement

Government security agencies can be quite effective after a public event or when they have criminal suspects. The greater challenge facing these organizations is to notice and react to warnings and indications prior to an event—for example, to discover a criminal's or terrorist's plot in time to intervene. Toward this goal, these organizations gather huge amounts of content from public and classified sources.

In the past, security agencies required analysts to examine documents manually to explore or substantiate theo-

ries. Increasingly, they use automatic extraction technologies to identify entities and link them to one another and to places, times, and events. Important entity types in this sector include people, places, organizations, weapons, chemical compounds, phone numbers, license plates, vehicles, and so on. Facts or events involving these entities might include purchase or sale events that connect people and organizations or that associate particular things or identification numbers with particular people.

Beyond lifting the burden from personnel, automatic extraction generally lets these organizations better direct their attention. Analysts can move beyond the conventional search to look for specific facts or types of occurrences and contextualize these results against the background of the content collection. With extracted information stored as structured databases, analysts can explore facts and relationships directly to look for significant patterns, trends, or anomalies. Ultimately, security organizations can apply statistical and mining techniques to automatically trigger human involvement. These possibilities illustrate various methods for mixing human attention and automation as well as theory- and content-driven approaches.

Pharmaceutical Research and Development

Improving the efficiency and effectiveness of the drug development process is a major business priority for pharmaceutical and biotechnology companies. Drug development typically spans 10 to 15 years and involves contributions from geneticists, biologists, chemists, chemical engineers, medical researchers, and other specialists. The current estimated cost of successfully bringing a new drug to approval is more than \$600 million. Revenue from a major drug, however, typically runs into billions of dollars.

Existing content—including scientific literature and the pharmaceutical company's internal content—contains information that can help improve time to market, decrease costs of unproductive efforts, and facilitate valuable discoveries. To support reuse and sharing of internal content, many life sciences companies have deployed search engines in their content management or portal efforts. As they have begun to see the limits of traditional search, they have started to deploy categorization solutions with specialized taxonomies or vocabularies—for example, MeSH (<http://www.nlm.nih.gov/mesh/meshhome.html>) or the Gene Ontology (<http://www.geneontology.org>).

Content mining in early stages of drug discovery can help identify the most promising avenues or cancel work on paths unlikely to succeed. Both external and internal documents often contain specific information about compounds, genes, proteins, diseases, and symptoms and establish links among these entities. A drug development team might use these patterns and statistics, for example, to understand diseases and mechanisms, to identify promising targets, and to optimize leads. A form of high-throughput information screening can isolate relevant connections between compounds and proteins, genes, diseases, and so on.

Content mining can also support strategic decision making in life sciences organizations. For example, manage-

ment must select therapeutic areas to invest in, understand the competitive situation in the marketplace, and manage effective patenting, partnering, and licensing strategies. Broader clinical, manufacturing, and marketplace activities all produce volumes of textual content including patents, adverse event reports, competitive analyses, analyst reports, industry news, clinical reports, and internal memos. Pharmaceutical companies can mine all of this content to support decision making and planning.

The library and information sciences discipline has long appreciated how important organization is to providing access to information. Without a disciplined cataloging system, finding and accessing relevant books become all but impossible. Similarly, your organization's content management system might be storing away documents never to be used again. But current commercial content analysis technologies can provide metadata about collections and documents, enabling applications for better access, routing, and mining of your company's precious information. ■

Ramana Rao is the CTO and founder of Inxight Software and the editor of the monthly newsletter Information Flow (<http://www.ramanarao.com/informationflow/>). Contact him at rao@inxight.com.

ADVERTISER / PRODUCT INDEX

NOVEMBER / DECEMBER 2003

Advertiser / Product	Page Number	Advertising Personnel	
Addison-Wesley	57	Marion Delaney IEEE Media, Advertising Director Phone: +1 212 419 7766 Fax: +1 212 419 7589 Email: md.ieeemedia@ieee.org	Sandy Brown IEEE Computer Society, Business Development Manager Phone: +1 714 821 8380 Fax: +1 714 821 4010 Email: sb.ieeemedia@ieee.org
John Wiley & Sons	57, 58	Marian Anderson Advertising Coordinator Phone: +1 714 821 8380 Fax: +1 714 821 4010 Email: manderson@computer.org	
McGraw-Hill	59		
Meghan Kiffer Press	58		
Advertising Sales Representatives			
Mid Atlantic (product/recruitment) Dawn Becker Phone: +1 732 772 0160 Fax: +1 732 772 0161 Email: db.ieeemedia@ieee.org	Midwest (product) Dave Jones Phone: +1 708 442 5633 Fax: +1 708 442 7620 Email: dj.ieeemedia@ieee.org	Midwest/Southwest (recruitment) Darcy Giovingo Phone: +1 847 498-4520 Fax: +1 847 498-5911 Email: dg.ieeemedia@ieee.org	Northwest/Southern CA (recruitment) Tim Matteson Phone: +1 310 836 4064 Fax: +1 310 836 4067 Email: tm.ieeemedia@ieee.org
New England (product) Jody Estabrook Phone: +1 978 244 0192 Fax: +1 978 244 0103 Email: je.ieeemedia@ieee.org	Will Hamilton Phone: +1 269 381 2156 Fax: +1 269 381 2556 Email: wh.ieeemedia@ieee.org	Southwest (product) Bill Wageneck Phone: +1 972 423 5507 Fax: +1 972 423 6858 Email: bill.wageneck@wageneckassociates.com	Japan German Tajiri Phone: +81 42 501 9551 Fax: +81 42 501 9552 Email: gt.ieeemedia@ieee.org
New England (recruitment) Barbara Lynch Phone: +1 401 739-7798 Fax: +1 401 739 7970 Email: bl.ieeemedia@ieee.org	Joe DiNardo Phone: +1 440 248 2456 Fax: +1 440 248 2594 Email: jd.ieeemedia@ieee.org	Northwest (product) Peter D. Scott Phone: +1 415 421-7950 Fax: +1 415 398-4156 Email: peterd@pscottassoc.com	Europe (product) Hilary Turnbull Phone: +44 1875 825700 Fax: +44 1875 825701 Email: impress@impressmedia.com
Connecticut (product) Stan Greenfield Phone: +1 203 938 2418 Fax: +1 203 938 3211 Email: greenco@optonline.net	Southeast (product/recruitment) C. William Bentz III Email: bb.ieeemedia@ieee.org Gregory Maddock Email: gm.ieeemedia@ieee.org Jana Smith Email: jsmith@bmmatlanta.com Phone: +1 404 256 3800 Fax: +1 404 255 7942	Southern CA (product) Marshall Rubin Phone: +1 818 888 2407 Fax: +1 818 888 4907 Email: mr.ieeemedia@ieee.org	Europe (recruitment) Penny Lee Phone: +20 7405 7577 Fax: +20 7405 7506 Email: reception@essentialmedia.co.uk